Project Report: Exploratory Data Analysis - Chicago Public Schools Dataset

Disclaimer

The following project report was written to satisfy the requirements of the final project for the Coursera Exploratory Data Analysis for Machine Learning course by IBM. It closely follows the materials outlined in the course including worked examples.

My code to produce figures as well as the figures are available in my github repository.

Objectives

For this exercise, I am training several linear regression models. The main purpose of this exercise is to see whether some features in a dataset (see description below) are able to **explain** the behavior of medical charges with respect to features in the dataset.

Description of Data Set

The dataset used in this work is the Chicago Public Schools Progress Report for the 2015/2016 school year. It provides information about all public schools in Chicago and also provides key metrics about school quality such as test score performance, attendance, and satisfaction survey data.

The dataset was downloaded from the City of Chicago Open Data Portal.

The data set is a CSV file with 153 columns containing information for 670 schools. For each school 153 parameters are being recorded. Some features are numeric, while other features are categorical. There are also features that are plain text, which will not be included in the analysis. Since the dataset includes all public schools in Chicago, there are several features that only apply to certain school times (example: grade specific test results).

The data types of the variables contained in the dataset are

Data Type	Counts
Numeric (Float)	81
Object	69
Numeric (Integer)	3

The majority of numeric variables consist of percentages for scores and or year numbers, while objects are often descriptive categories giving additional information.

Initial Plan for Data Exploration

To get an initial overview about the data, I will select a subset of columns relating to school performance that have numerical or categorical values and will review descriptive statistics such as mean, range, median as well as value counts for categorical data. I will also assess whether there is missing data and treat missing data accordingly by either elimination of the rows or gap filling with mean values. I will also create correlation plots between variables using seaborn.

Actions Taken for data cleaning and feature engineering

have selected a subset of numeric and categorical features including:

- school type
- attainment rating
- culture climate rating
- survey response rate
- healthy school certification
- suspension rate
- student attendance
- teacher attendance
- dropout rate
- freshman on track track

These data columns were also renamed to facilitate the analysis. I have not removed data with missing values (since for example dropout rate does not apply to elementary schools), but this should be done at a later stage depending on the further desired analysis.

Figure 1 shows a pairwise plot of the chosen data columns separated by *student attainment rating* which may be considered a potential target variable for statistical modeling.

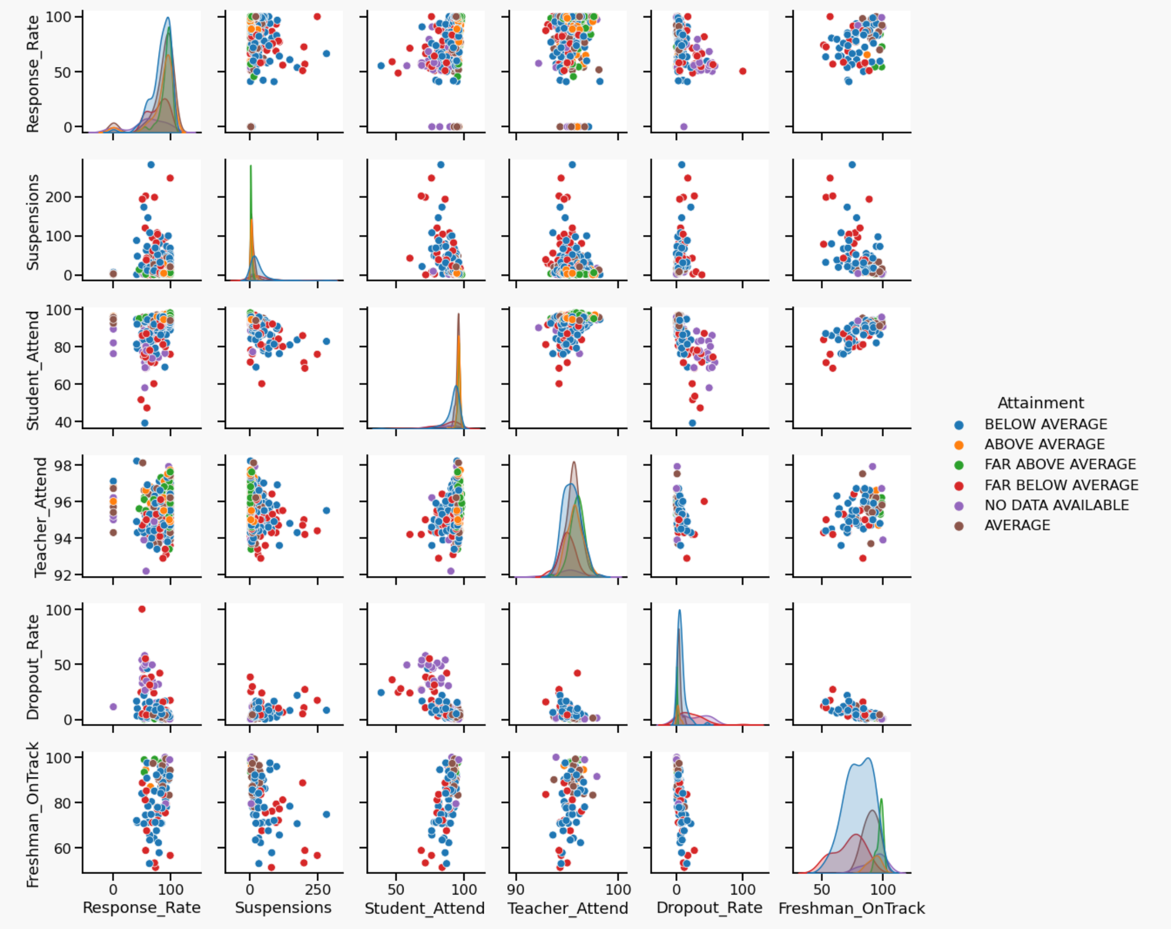
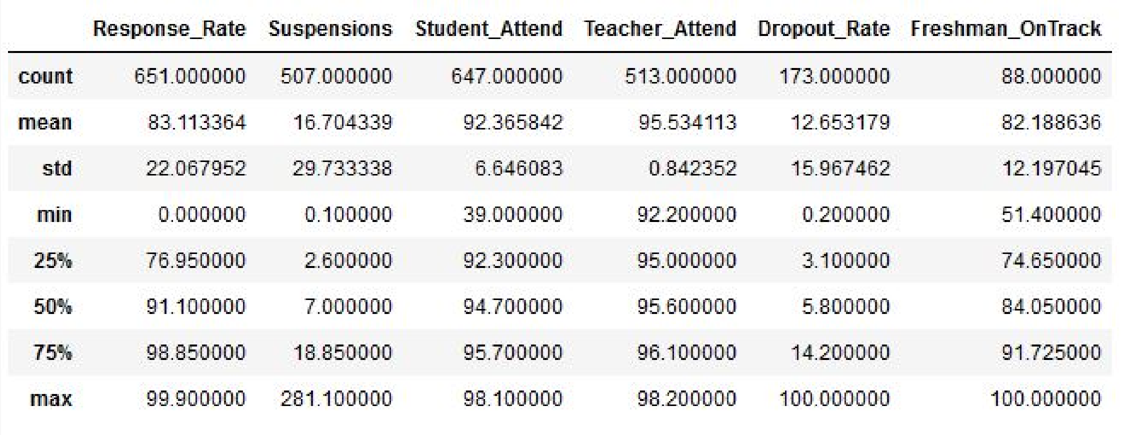


Figure 1: *Pairwise plot of data columns separated by student attainment*

I calculated descriptive statistics for numerical columns (Figure 2):



	Response_Rate	Suspensions	Student_Attend	Teacher_Attend	Dropout_Rate	Freshman_OnTrack
count	651.000000	507.000000	647.000000	513.000000	173.000000	88.000000
mean	83.113364	16.704339	92.365842	95.534113	12.653179	82.188636
std	22.067952	29.733338	6.646083	0.842352	15.967462	12.197045
min	0.000000	0.100000	39.000000	92.200000	0.200000	51.400000
25%	76.950000	2.600000	92.300000	95.000000	3.100000	74.650000
50%	91.100000	7.000000	94.700000	95.600000	5.800000	84.050000
75%	98.850000	18.850000	95.700000	96.100000	14.200000	91.725000
max	99.900000	281.100000	98.100000	98.200000	100.000000	100.000000

Figure 2: *Pairwise plot of data columns separated by student attainment*

Given that student attainment is a likely target variable for modeling it makes sense to also have a look at how this variable is distributed:

Student Attainment	**Counts**
FAR ABOVE AVERAGE	89
ABOVE AVERAGE	110
AVERAGE	157
BELOW AVERAGE	183
FAR BELOW AVERAGE	73
NO DATA AVAILABLE	54

The following actions were performed regarding feature engineering:

1. categorical data (healthy school certification, attainment rating, school type) were encoded to numeric data using one-hot-encoding.
2. columns with high skew (Suspensions, Dropout_Rate, Response_Rate, Student_Attend) were log transformed to improve normality (Figure 3).

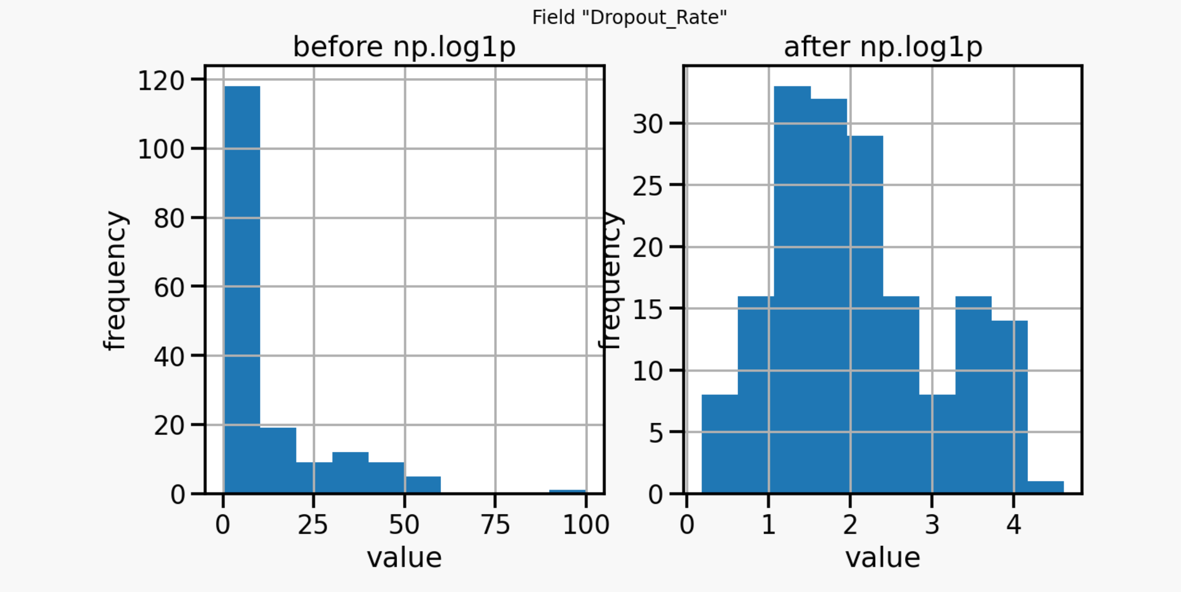


Figure 3: *Example of result for applying logarithmic transform to a highly skewed column: left: before; right: after transform*

Key Findings and Insights

Initial analysis of the data (see Figure 1) indicates the following:

- Schools with attainment scores of 'below average' and 'far below average' have considerable higher 'dropout rates' and lower rates of 'students on track'
- There is a strong correlation between student attendance and *on track status* highlighting the fact that going to class is associated with higher student achievement, even though confounding factors may be at play.
- There is also negative correlation between *student on track* status and *dropout rates*
- Suspensions are rare at higher performing schools and there is considerable variation between suspension rates even at lower performing schools.
- There seems to be little variation in teacher attendance between different performing schools.

Hypothesis formulation

Hypothesis 1

H_1: There is a significant relationship between student attendance and student on track status.

$H_{0,1}$: There is no significant relationship between student attendance and student on track status

Hypothesis 2

H_2: Schools with higher attainment rating have better student outcomes as evidence by on track status

$H_{0,2}$: There is no significant difference between schools with different attainment levels

Hypothesis 3

H_3: There is a significant relationship between student attendance and teacher attendance.

$H_{0,3}$: There is a significant relationship between student attendance and teacher attendance

I proceed to evaluate **Hypothesis 2** by comparing 'above average' and 'below average' schools. For this I do a t-test to identify whether both distributions have a significantly different mean for 'on track' status

Hypothesis testing and interpretation of results

The conducted t-test between schools with above and below average attainment scores yield a **p-value of 0.02**, leading to rejection of the Null-hypothesis is being rejected. Therefore, the test shows that indeed, schools with higher attainment rating have significantly different dropout rates from schools with low attainment.

Suggestions for next steps

It would be interesting to perform further explanatory data analysis, e.g. investigate which factors intrinsic and extrinsic to schools contribute to successful school completion. To do so, one could fit a machine learning model to the data and then investigate which coefficients/ features of the model have the highest explanatory power.

Additional Data Needed

The CPS school scorecard dataset only includes data that is collected within the school system. However educational outcomes are also highly dependent on socioeconomic and geographic factors. Therefore it would be good to merge the CPS dataset with other data sources such as census data or data about Chicago neighborhoods in order to better understand which factors affect school quality and student success.